Data Science: The Impact of Machine Learning

Shidul Islam
Department of Computer Science and Engineering
Daffodil International University
Dhaka 1207
Bangladesh
Email: rahaddiu@gmail.com

Abstract: In this paper, I prove my promise that Machine Learning is one of the most important parts to provide tools and methods to go deeper and nurture the data properly. The most amazing part is to analyze the large chunks of data in a very precise way, and high-value predictions that can guide better decisions and smart actions in real-time without human intervention. I give an overview over different proposed structures of Data Science and mention the impact of Machine learning such as algorithms, model evaluation and selection, pipeline. I also indicate all misconceptions when neglecting Machine learning reasoning.

Keywords: Structure of Data Science, Impact of Machine learning on Data science, Misconception in Data science

1. Introduction

Data Science such a field where uses processes, scientific method, algorithms, and systems to gain knowledge from a huge structural and unstructured data [1][2]. Data Science is combined with statistics, data analysis, machine learning, and their related methods for analyzing and understand the actual phenomena with data [3].

In 1962, John Tukey described a field and that was called "data analysis", which is today's modern data science [4].

Data science is a field that extracts the knowledge from datasets, and that datasets typically large [5]. The field encompasses analysis, preparing data for analysis, and presenting findings to inform high-level decisions in an organization. As such, it incorporates skills from computer science, mathematics, statistics, information visualization, graphic design, and business [6][7]. In 2015, the American Statistical Association identified database management, statistics and machine learning, and distributed and parallel systems as the three emerging foundational professional communities [8].

At present, Machine learning and data science are working hand in hand. Take into consideration the definition of machine learning - the capability of a machine to generalize knowledge from data. Without data, machines can't learn and Machine learning is only as good as the data it is given and the ability of algorithms to consume it. Going forward, basic levels of machine learning will become an urgent need for every data scientist [9].

In my view, machine learning methods and tools are crucial in most fundamental steps in Data Science. Therefore, the premise of my contribution is:

Machine learning is one of the most important parts to provide tools and methods to go deeper and nurture the data. This paper mainly addresses the major impact of machine learning on the most important steps in Data Science.

2. Steps In Data Science

Data Science is a process and that process not to say it's mechanical and void of creativity. But, when anyone digging into the stages of processing data, from munging data sources and data cleansing to machine learning and eventually visualization. Unique steps are involved in transforming raw data into insight [10].

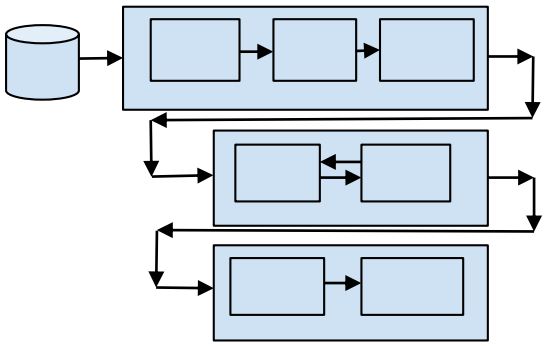

Figure 1. The data science pipeline

In my view, the main steps in Data Science have been inspired by CRISP-DM. Data Science is a sequence of the following steps: Data Acquisition and Enrichment, Data Storage and Access, Data Exploration, Data Analysis and Modeling, Optimization of Algorithms, Model Evaluation and Selection, Representation and Reporting of Results, and Business Deployment of Results.

Usually, these steps are not just combined once but are iterated in a cyclic loop. Here, two or more steps are common and this holds especially for the steps optimization of Algorithms, Model Evaluation and Selection.

From (Figure 1), we are seeing that Algorithms and Model Evaluation and Selection are part of Machine learning. I will highlight the role of Machine learning, where it is heavily involved.

2.1 Algorithms

Machine learning algorithms are programs (math and logic) that adjust themselves to perform better as they are exposed to more data and the "learning" part of machine learning means that those programs change how they process data, much as humans change how they process data by learning. So, a machine learning algorithm is a program with a specific way of adjusting its own parameters, given feedback on its previous performance making predictions about a dataset [11].

There are mainly four types of machine learning algorithms: supervised, semi-supervised, unsupervised, and reinforcement [12].

But, inspired by Reena Shaw from KDnuggets, the most important machine learning algorithms for data science [19]:

- Supervised learning.
- Unsupervised learning.
- Ensembling learning.

2.1.1 Supervised learning

Supervised learning is where anyone has input variables (A) and an output variable (B) and uses an algorithm to learn the mapping function from the input to the output[13].

$$B = f(A)$$

The goal is to approximate the mapping function so well that when anyone has new input data (A) that can predict the output variables (B) for the data [13].

Types of supervised learning algorithms include Active learning, classification and regression [14].

2.1.1.1 Classification is a task that requires the use of machine learning algorithms that learn how to assign a class label to examples from the problem domain. Like, identify the emails as "spam" or "not spam". There are perhaps four main types of classification [15]. But I will discuss it in line with data science.

a) **Binary or binomial classification** is the task which classifies the elements of a given set into two groups (predicting which group each one belongs to) on the basis of a classification rule. Contexts requiring a decision as to whether or not an item has some specified characteristic, or some typical binary classification include [16]:
   - Model testing to determine if a patient has a certain disease or not - the classification property is the presence of the disease.
   - A "pass or fail" test method or quality control in factories, i.e. deciding if a specification has or has not been met - a Go/not go classification.

   Popular algorithms for binary classification include:
   - Logistic Regression.
   - K-Nearest Neighbors.
   - Decision Trees.

b) **Multiclass or multinomial classification** is the problem of classifying instances into one of three or more classes. (classifying instances into one of two classes is called binary classification) [17].
   The existing multi-class classification techniques can be used categories into (i) Transformation of binary (ii) Extension from binary (iii) Hierarchical classification [18].
   Transformation to binary discusses strategies for reducing the problem of multiclass classification to multiple binary classification problems. It can be categorized into One vs Rest and One vs One. The techniques developed based on reducing the multi-class problem into multiple binary problems are called problem transformation techniques [17].
   - One-vs-Rest: Fit one binary classification model for each class vs. all other classes [15].
   - One-vs-One: Fit one binary classification model for each pair of classes [15].

   Extension from binary section is a technique that extending the existing binary classifiers to solve multi-class classification problems [17].
   Hierarchical classification is the multi-class classification problem by dividing the output space i.e. into a tree. Each parent node is divided into multiple child nodes and the process is continued until each child node represents only one class [17].
   Popular algorithms that can be used for multiclass classification include:
   - K-Nearest Neighbors.
   - Decision Trees.
   - Naive Bayes.
   - Random Forest.

   Algorithms that are designed for binary classification can be adopted for multi-class problems.

2.1.1.2 Regression is a Machine learning algorithm that is used to predict a continuous value. Predicting prices of a house given the features of the house like size, price, etc [20]. There are two commonly used variations of regression [21].

a) **Linear regression** is the most widely known modeling technique and that is usually the first few topics that people pick while learning predictive modeling.
   Linear regression establishes a relationship between the **dependent variable(Y)** and one or more **independent variable(X)** using a best fit straight line(also known as a regression line) [22].
   It is represented by an equation **Y = b + mx + e**, where b is the intercept, m is the slope of the line, and e is error term [22].

b) **Logistic regression** is a supervised learning classification algorithm that is used to predict the probability of a target variable. The nature of the target or dependent variable is dichotomous, which means there would be only two possible classes [23].
   Logistic function (also called **sigmoid function**) is an S-shaped curve that maps any real-valued number to a value between 0 and 1 [24].

   $$1 / 1 + e^{-z}$$

   Here, **e** is the Euler number, and z is a boundary function.

   $Z = ax + b$, where x is an input variable, a is coefficient, and b is biased.
   Here, using the sigmoid function to make a prediction.

   $$Y = 1 / 1 + e^{-z}$$

   Or, $Y = 1 / 1 + e^{-(ax + b)}$ [24]

2.1.2 Unsupervised learning

Unsupervised learning is a machine learning algorithm that is used to draw inferences from datasets consisting of input data without labeled responses [25].

Unsupervised learning problems can be further grouped into clustering and association problems [26].

a) **Clustering** is a machine learning technique that is involved in the grouping of data points. Given a set of data points and using a clustering algorithm to classify each data point into a specific group [27].
b) An **Association** rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y [26].

Some popular examples of unsupervised learning algorithms for data science [26]:

- K-means for clustering problems.
- Apriori algorithm for association rule learning problems.
- principal component analysis (PCA).

2.1.3 Ensemble learning

Ensemble means to unify the results of more than one learner for improved results, by voting or averaging. Basically, voting is used for the time of classification, and averaging is used during regression. The idea is that ensembles of learners perform better than single learners [28].

There are three types of ensembling algorithms [28]:

- Bagging.
- Boosting.
- Stacking.

2.2 Model evaluation and selection

A machine learning model can be a mathematical representation of a real-world process [29], and Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data as well as how well the chosen model will work in the future also. Evaluating model performance with the data which is used for training is not acceptable in Data Science because it can easily generate overfitted models. Evaluating models there are two methods in Data Science, Hold-Out, and Cross-Validation. To avoid overfitting, both methods use a test set (not seen by the model) to evaluate model performance [30].

2.2.1 Hold-Out

In this method, the mostly large dataset is randomly divided into three subsets [30]:

a) The **Training set** is a subset of the dataset used to build predictive models.
b) The **Validation set** is a subset of the dataset that is used to assess the performance of the model built in the training phase. It provides the best platform for selecting the best performing model and not all modeling algorithms need a validation set.
c) A **Test set** is a subset of the dataset to assess the likely future performance of a model. If a model fit to the training set much better than it fits the best set, overfitting is probably the cause.

2.2.2 Cross-Validation

When a limited amount of data is available, to achieve an unbiased estimate of the model performance we use K-fold cross-validation. In K-fold cross-validation, we can divide the data into k subsets of equal size and build models K times, each time leaving out one of the subsets from training and use it as the test set. If K equals the sample size, this is called "leave-one-out" [30].

Model evaluation can be divided into two sections [30]:

- Classification Evaluation.
- Regression Evaluation.

Why we need the distinction between model selection and model evaluation? The reason is overfitting. Basically, model evaluation is used to estimate the generalization error of the selected model, i.e., how well the selected model performs on unseen data [31].

2.3 Pipeline

A machine learning pipeline that is used to help automate machine learning workflows. It is a sequence of data to be transformed and correlated together in a model and then be tested and evaluated to achieve the result, whether positive or negative [32].

Machine learning pipelines consist of several steps to train a model [32].

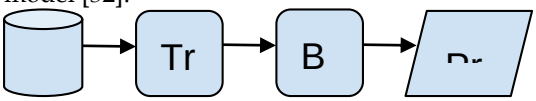

Figure 2. The machine learning Pipeline [33]

3. Misconception

One huge task in data science entails spending a lot of time in finding sound datasets and formulating suitable scientific questions. Data Scientists usually take a lot of time to make sure that they have the right data at hand for them to generate good results for their algorithms. It is sad to note that most people perceive data science as an activity that involves 10 percent data extracting and cleaning and 90 percent modeling [34].

Invariably, the request for an ad-hoc piece of analytics work pre-supposes the availability of data that forms the basis for this work. This may seem like the most basic kind of assumption, but it wouldn't be too rare to be asked to analyze some dataset, only to realize that it's non-existent, inaccessible, lacking common identifiers, or aggregated/summarised at too high a level [35].

Data scientists always are happy to get nice structured format datasets but the harsh truth is that one can only find those datasets if a data engineer or data scientist designed them [34].

In Data Science, past performance implies Future results! Modeling assumptions can be held as absolute truths after experiments and variables are normally distributed unless otherwise specified [36].

4. Conclusion

Following the above assessment of the abilities and the impact of machine learning my conclusion is:

The importance of Machine learning in Data science is not negligible and also in computer science. In particular,

algorithms and modeling is too significant for high-value prediction.

Interestingly, most of the machine learning engineers are working in data science. Some job sector they first want their previous machine learning experience for data science field.

In conclusion, machine learning engineers can play a vital role in modern Data Science.

References

1. Dhar, V. (2013, December). "Data Science and Prediction." Communications of the ACM, 56(12), 64-73.
2. Leek, J. (2013, December 12). "The keyword in Data Science is not Data, it is Science." Simply Statistics.
3. Hayashi, C., Yajima, K., Bock, H., Ohsumi, N., Tanaka, Y., & Baba, Y. (1996). "Data Science, Classification, and Related Methods." N.p.: springer.
4. Donoho, D. (2015). "50 years of Data Science."
5. Data Science Association. (2020). "About Data Science." In . (Ed.).
6. O'Neil, C., & Schutt, R. (2013). "Doing Data Science." N.p.: O'Reilly Media, Inc.
7. Driscoll, M. E. (2009, May 27). "The three sexy skills of data geeks." m.e.driscoll: data utopian.
8. "ASA Statement on the Role of Statistics in Data Science." (2015, October). AMSTATNEWS.
9. Tavasoli, S. (2020, January). "The Importance of Machine Learning for Data Scientists." Simplilearn.
10. Jones, M. T. (2018, February 1). "Data, structure, and the data science pipeline." IBM.
11. Machine Learning Algorithms Some Basic Machine Learning Algorithms. Pathmind.
12. Wakefield, K. "A guide to machine learning algorithms and their applications." SAS.
13. Brownlee, J. (2016, March 16). "Supervised and Unsupervised Machine Learning Algorithms." Machine Learning Mastery.
14. Alpaydin, E. (2014). "Introduction to Machine Learning (3rd ed.)." N.p.: MIT Press.
15. Brownlee, J. (2010, April 8). "4 Types of Classification Tasks in Machine Learning." Machine Learning Mastery.
16. "Binary classification." (2011, May). Wikipedia.
17. "Multiclass classification". wikipedia.
18. Mohamed, Aly (2005). "Survey on multiclass classification methods" (PDF). Technical Report, Caltech.
19. Shaw, R. (2017, October). "Top 10 Machine Learning Algorithms for Beginners." KDnuggets.
20. Dave, A. (2018, December 4). "Regression in Machine Learning." Data Driven Investor.
21. GOEL, A. (2018, June 13). "What Is a Regression Model"? Magoosh.
22. Ray, S. (2015, August 14). "7 Regression Techniques you should know!" Analytics Vidhya.
23. "Machine Learning - Logistic Regression". Tutorialspoint.
24. "Learn Logistic Regression using Excel – Machine Learning Algorithm." (2017, December 26). New Tech Dojo.
25. "Unsupervised Learning." MathWorks.
26. Brownlee, J. (2019, August 12). "Supervised and Unsupervised Machine Learning Algorithms." Machine Learning Mastery.
27. Seif, G. (2018, February 5). "The 5 Clustering Algorithms Data Scientists Need to Know." Towards Data Science.
28. Shaw, R. "The 10 Best Machine Learning Algorithms for Data Science Beginners." Dataquest.
29. Bhattacharjee, j. (2017, October 28). "Some Key Machine Learning Definitions." Technology-Nineleaps.
30. Sayad, D. S. "Model Evaluation." saedsayad.
31. Schönleber, D. (2018, December 10). "A "short" introduction to model selection." Towards Data Science.
32. M, S. (2019, December 11). "What is a Pipeline in Machine Learning? How to create one?" Analytics Vidhya.
33. "Data Science Full Course - Learn Data Science in 10 Hours | Data Science For Beginners." (2019, August 18). Edureka.
34. SHANIN, S. (2018, April 10). "6 Common Data Science Fallacies You Should Avoid." eteam.io. Retrieved from https://www.eteam.io/
35. Brennan, S. (2017, September 17). "The Ten Fallacies of Data Science." Towards Data Science.
36. Masood, A. (2016, May 25). "The Fallacies of Data Science." Data Science Central.